CLOUD COMPUTING AND IMPLICATIONS FOR PUBLIC POLICY

BRIE WORKING PAPER 194

[PREVIEW]

December, 2010

Kenji E. Kushida, Jonathan Murray, John Zysman

Introduction: Cloud Computing, the Next Computing Platform

It has been a truism of the computing industry that every ten to fifteen years we see the emergence of a new platform that disrupts previous models of computational service delivery. Each new platform combines changes in hardware, software architecture and service delivery infrastructure to offer more flexible and efficient methods for delivering existing services and the opportunity to create a new generation of entirely novel services which further extends the opportunity for economic value creation. Each new platform brings with it the opportunity to re-craft, work, processes, businesses structures and even entire markets. These new platforms have spawned entirely new ecosystems of opportunity with entrepreneurs and firms finding innovate new ways to exploit these new computational resources. These changes were witnessed with the introduction of the mainframe computing model in the nineteen sixties, the mini-computer in the seventies, the personal computer in the nineties and the introduction of internet computing models which really took off at the start of the 21st century. None of these new platforms truly replaced previous computational models; they have invariably built upon the models of the past while establishing new capabilities in their own right.

The new emerging computing platform is the set of technologies and service delivery models collectively known as “Cloud Computing.” Perhaps more so than its predecessor platforms, Cloud Computing builds upon previous platforms while at the same time offering a radically new approach to the delivery of computational resources.

Cloud Computing is, in its essence, information technology services delivered as an “enhanced utility,” with a set of characteristics that closely parallel that of other utilities such as electricity, gas, and water – though with some important differences, as we will see later. With utilities, the resources are always available, and we can generally consume as much as we like – as long as we can pay for it. We consume the service through standard connections and devices,
and we do not need to know or care how the service is run on the back end (such as a power plant or water processing plant). Moreover, we are free to do what we like with the resources as long as we abide by the law, and the utility provider is indifferent whether the electrons they serve you are used to run your television, vacuum cleaner or personal computer. Similarly, with Cloud computing, you contract with a cloud provider to provide various information technology services: disk storage, computational processing, network bandwidth or a platform on which to run your accounting or customer relationship management system.

Unlike a traditional utility – which must make best efforts to provide uninterrupted supply - a Cloud provider will generally offer a range of Services Level Agreements (SLAs) which, guarantee differing levels of service availability depending on price. Once you agree to a particular service level you are then free to decide how much of each resource is needed for your specific requirements. These needs can change over time, and the service dynamically adjusts to meet those changing needs. In the end you do not care about the technical aspects of how the provider runs their services, and the service provider does not care what ‘workload’ you are running on their system. Nor do they care about the device on which you consume the service. Ultimately, you can consume any amount of the service you need at any time, limited only by the price you are able or willing to pay.2

A paradigmatic example of a Cloud service acting as an “enhanced utility” can be witnessed in new models for rendering animation graphics for movies. The traditional approach required each production house to create its own – very expensive – ‘Render Farm’ consisting of hundreds or thousands of computers dedicated to converting the artists’ digital drawings into the thousands of individual frames which create the moving image. A new approach being pioneered by Pixar, which is currently beta testing a new Cloud based approach, which offers its powerful Renderman software to third party production studios. The service is accessible through a web based interface which provides a very simple method for letting the user trade off cost vs. completion time for each job. For ‘Rush’ jobs the Cloud service dynamically allocates the required computing resources to complete the job in the required time. This results in a much higher charge for the user. If the user is working to a smaller budget then the jobs take longer as the system allocates a lower amount of computational resources.

The transition to Cloud computing is likely to occur faster than previous platform transitions. Previous new platforms’ adoption rates were often limited by the need to make substantial capital investments, the time to train, re-train and acquire the technical expertise to exploit the platform, and the cost of re-writing existing bodies of application software. The new cloud-computing model, however, eliminates almost all of these traditional barriers to adoption.

A major part of the transformative potential of Cloud computing lies in its ability to open the door to entirely new and innovative business models. Cloud computing dramatically lowers the very significant barriers to entry for new players in a broad array of markets. In the past, delivering new information technology based services required substantial capital outlays to build the ICT capabilities. Delivering the services also incurred ongoing operational costs for running the information technology organizations. The new cloud-computing model, however, removes these costs almost entirely. New players can now contract with a cloud-computing provider to deliver the required service on a pay-as-you-go basis, and the firm does not need its own full service information technology staff or physical computing infrastructure. These

---

2 Put more formally, these are the characteristics of scalability, illusion of infinite resources, dynamic allocation of resources.
advantages will have the most profound effect on startups and small to mid-sized enterprises – critical sources of dynamism in any robust and growing economy. For large enterprises with significant internal datacenters, an application of Cloud computing architectures can increase the efficiency and flexibility of existing computing resources. Perhaps more importantly, Cloud computing can augment the enterprises’ internal resources by providing opportunities for groups or division to experiment with tasks or services that require massive amounts of computing resources for a short period of time – for example, running an experiment that benefits from the equivalent of substituting 10 servers for 1000 hours with 1000 servers for 10 hours.

For societies and economies, Cloud computing clearly provides opportunities. However, it also raises significant questions about the potential risks. Since the large, global-scale Cloud computing providers are building out massive infrastructures with processing and data storage taking place across many different countries, interconnected into a seamless whole by massive networks. The benefit of accessing always on, highly reliable, and massively scalable services, is significant as is the opportunity cost of not having access, but in most countries, the source of these services will lie outside their national boundaries. What happens in the unlikely event of a catastrophic failure of service, or a provider going out of business? How do questions of legal jurisdiction become resolved in a manner fast enough that the economic potential of Cloud services can be harnessed in the national economy? Politically, since Cloud computing inherently deals with information, issues of privacy and security quickly become paramount. Old political settlements can be reopened, while new issues, constituencies, and interest groups can be reconfigured and mobilized in new ways. From a regulatory perspective, another key issue concerns the interoperability of data between services, and the degree to which firms or consumers might avoid being locked in to one vendor’s services.

This paper cuts into Cloud computing by asking and answering three fundamental questions. First, what exactly is Cloud computing, and how do we make sense of the rapidly developing market with a diverse array of service offerings labeled “Cloud”? Part I provides a brief background and useful set of conceptual tools to understand the rapidly developing Cloud computing market, cutting through much of the confusion that inevitably occurs when a wide variety of firms begin branding their products and services as “Cloud.”

Second, what would a corporate executive need to know about Cloud computing? A focus on what CEOs are interested in helps us cut through the technical and architectural details to give us a vantage on how the business of Cloud computing will be adopted by large enterprises. Part II provides such a vantage. While some CEOs have considered IT as a “necessary evil” – a cost to be controlled and risk to be managed – others have viewed IT as the most strategic of all capabilities and the lifeblood upon which the success of the firms depends. Cloud computing provides opportunities for both cost-cutting focused and strategically oriented CEOs.

Third, what are the policy and regulatory issues that can affect how Cloud computing diffuses throughout the world? Policymakers and regulators face a stark reality: the rate of global adoption of these new services and their potential scale and impact means you cannot sit idly by while de facto reality becomes the rules by which everyone operates. At the same time, national rules and regulations will matter in how Cloud services develop across the globe. Part III presents several critical policy issues and emerging debates that will affect the development of Cloud services.

Finally, for any coherent and comprehensive discussion of Cloud computing, four different perspectives must be considered. Firms which contract for the provision of cloud-
computing services to enable their own businesses (firm users); Consumers who ultimately utilize services provided by firms utilizing a cloud-computing model (consumer users); Cloud service providers themselves who are responsible for building out and managing the infrastructure necessary to deliver this new computing model, and finally; the policy makers and regulators in each market who must understand and keep pace of the impact cloud-computing will have on market governance.

Part I: What is Cloud Computing Anyway, and How Can We Understand the Rapidly Developing Market?

Cloud computing is rapidly emerging as the next computing platform. Just as in the case with all technological platforms in their infancy, but perhaps more so due to the rapid emergence of this platform, the term “Cloud computing” is surrounded by a great deal of confusion. The confusion is exacerbated because all manner of firms are attaching the term to their service offerings as a marketing move. So what exactly is Cloud computing, and what is new about it? (And how is it different from conventional online services, or the “Web 2.0” that was commonly heard until recently?) What is the nature of competition (i.e. what are firms competing about?) and what is the best way to understand how the market is unfolding? Who are the major players, what have they been doing, and what are their strategies? (Everybody, ranging from Microsoft to IBM, AT&T, and Salesforce.com seems to have a “Cloud solution/service” but what are they actually doing?) These are the questions addressed in Part I.

Cloud Computing: an Enhanced Utility and Competitive Proposition

First, what is Cloud Computing? We avoid an overly technical definition, focusing instead on the functionality it provides. Since the market is still developing, definitions are bound to change as new services and business models introduce new possibilities, but to cut through much of the confusion, our operating definition is as follows:

“The provision of computing infrastructure, platform or application services as a utility, which can be consumed by any Internet connected device, using open standard protocols. Variability in demand is satisfied through the dynamic and

---

3 Our operating definition builds upon definitions from National Institute of Standards and Technology (NIST) and an influential study from the University of California Berkeley. Cite NIST definition website.

4 The NIST definition begins with Cloud Computing providing “on-demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." See Appendix 1 for the full NIST definition.
automatic (Elastic) provisioning of pooled hardware, network, software, and service resources that are billed for on a pay-as-you-go basis.”

These characteristics, when taken together and combined in a service model, are significantly different from previous computing platforms.

To capture the nature of cloud computing, it is useful to distinguish between the perspectives of consumers – those that utilize the services built on cloud infrastructure, firms who leverage cloud infrastructure to deliver services to consumers, and providers, the companies whose business models include providing Cloud computing services.

For consumers the particular computing platform utilized by their favorite online store, internet video or social networking site is essentially immaterial. The consumer does not need to know or care how these services are created and delivered. However, without cloud-computing as a new platform, many of these innovative new services would not exist. In a very real sense the cloud approach has made it economically viable to create a new generation of services which would have been prohibitively expensive to create just five or ten years ago. For this reason cloud-computing has had a significant impact on the types of services consumers now have access to.

The situation for firms is entirely different. Whereas consumer do not need to know or care about how the services is delivered the firms which create and run these services, care very much. Cloud Computing is what we call an “enhanced utility.” It enables the low cost delivery of massive amounts of computational resources that would have been inconceivable even 10 years ago. The price point at which cloud-computing resources can be offered, and their global reach and scalability have catalyzed and entirely new generation of value creating service innovation. Former upfront capital investment in storage, network and computing resource has been eliminated. This means that a new service can be created with almost zero capital cost, with firms incurring only ongoing operational costs in direct proportion to the number of consumers using the service.

For all sizes of firms, from startups through large enterprises, as soon as large portions of their resources and energy can be diverted away from constructing and maintaining computing infrastructure, they are freer to innovate and experiment. (Just as in most places, companies do not have to concern themselves with constructing and maintaining the infrastructure to generate and deliver electricity or build dams and reservoirs to procure water supplies.) Cloud computing enables firms to establish leadership positions by leveraging the leading edge of technology since they do not have to worry about restructuring their datacenters and computing infrastructure as soon as new developments become available – just as users of utilities can benefit from improvements in electricity generation or smart grids. This essentially eliminates the risks previously associated with such strategies allowing for rapid innovation cycles and early market entry.

The range of Cloud choices expand for large enterprises or government agencies. These entities have significant existing investments in their own data centers and computing infrastructures. Cloud is often seen as additive to, rather than as a replacement for existing capability. Three common scenarios include: adding external cloud service capability to extend the utility of existing internally hosted services (Hybrid); re-architecting existing internal systems to implement Cloud-like scaling and dynamic resource allocation methodologies (Private); and the adoption of public cloud infrastructure to deliver new services as already
discussed (Public.) Most large enterprise and government agencies will end up implementing some version of each of these strategies.

For providers, **Cloud Computing is a competitive service opportunity.** The major providers of Cloud computing service have built massive global-scale datacenters in multiple locations around the world. They compete on the breadth, quality, reliability and cost of the services they deliver to firms and consumers. Providers make massive capital investments in computing resources to ensure they stay ahead of the demand curve. A typical global data center costs approximately $500 million, and Apple is reportedly investing $1 billion in its new North Carolina cloud data center. It is this investment that enables them to offer the illusion of infinite resource availability, allowing workloads to scale up and down rapidly. It is also the magnitude of the investment required which will necessarily limit the number of truly global cloud service providers. Very few technology firms have the capital resources needed to be able to invest such vast sums for building out cloud infrastructure.

The provider market segments according to the type and scale of service being offered. This segmentation enables smaller providers to find a competitive niche without necessarily incurring the capital investment costs required to be a global provider. Examples include those providing accounting, customer relationship management or other software applications as a service (SaaS), small scale providers of ‘Community’ or ‘Local’ cloud infrastructure to meet the needs of a geographically or functionally narrow set of consumers, and finally those providing various component cloud technologies which firms can configure to provide highly customized private cloud services.

**Understanding The Unfolding Cloud Computing Market**

Whenever a new technological, industry, or production paradigm takes hold, it is critical that we develop conceptual frameworks that are intuitive and useful to capture the fundamental industry dynamics at an early stage; these conceptions must go beyond a technological description of the architecture. Cloud Computing is exactly at such a stage, and there is still much confusion around the search for the most meaningful conceptions. A diverse range of firms offering a wide variety of services are all labeling themselves as offering “Cloud services,” leading to some unnecessary, but often intentional, levels of confusion.

We introduce two frameworks to sort through how different users, service providers, their strategies, and policy issues come into play. The first is a conception of a **service stack**, which enables us to understand how the various service functions such as Software as a Service (SaaS) and Platform as a Service (PaaS), which are provided from within the datacenters, depend on the networks through which these services are delivered, and the devices on which they are consumed. This will be particularly useful in understanding major policy issues that will affect the development of Cloud services in different national regulatory environments.

The second is a **typology of deployment patterns** by service providers. The set of dimensions we put forth will enable readers to place firms (and media articles) describing their services as “Cloud” into proper context. This will be useful to build an industry map, as well as to understand how various national markets and service providers are developing.
The Stacks Conception

Figure 1: The Cloud Computing Services Stack

Access Device Operating System
(for PCs, netbooks, smart phones, etc)

Access Device Hardware
(eg., PCs, netbooks, smart phones, etc)

Network Services
(eg., CDN/ADN Content Delivery Networks, Application Delivery Networks)

Network Infrastructure
(eg., DSL, FTTH, 3G, 4G, WiMax)

Applications, Content
(Software as a Service SaaS)

Platform (Software Environment)
(Platform as a Service PaaS)

Infrastructure Resources
(Infrastructure as a Service, IaaS)

Here we present a very simple explanation, leaving more fleshed out details to the Appendix. Cloud services offering include the functions provided from datacenters, access networks through which they are connected to users, and the access devices.

Within the Cloud data centers, the three layers of Applications (SaaS), Platform (PaaS), and Infrastructure Resources (IaaS) are best understood as different layers of abstraction on top of the underlying cloud service components. For example SaaS offerings, usually targeting firms and consumers, typically provide the user with access to powerful, domain specific applications such as accounting, customer relationship management or a health care information system. Users are confined to operate within the parameters set by the software, just as they are when using Microsoft Word on their own laptop. At this level of abstraction there is no need to know anything about the underlying cloud architecture being utilized to deliver the service. PaaS, provides developers with programmatic access to the underlying cloud infrastructure services without them having to know the precise details about how those services are implemented. PaaS enables developers to create arbitrary applications and services that take full advantage of the global scale, reliability and other attributes of the underlying cloud infrastructure. IaaS enables
firms’ IT staff and developers to move existing workloads – a web site, database application et cetera – to the cloud. IaaS replaces the firm’s own hardware and software IT infrastructure with a ‘Virtual’ version of the same environment run by the cloud provider. This approach has the ability to offer significant operational cost savings. Scaling typically remains limited by the original design and architecture of the application. To take full advantage of the cloud an application would need to be re-written to make use of PaaS services.

The Access Networks and Access Devices layers are useful in understanding many of the major Cloud providers’ strategies. Google and Microsoft, for example, have been buying significant amounts of physical network infrastructure around the world. (In fact, in 2009, Google ranked third worldwide in the total volume of Internet traffic carried over its networks, displacing the giant telecos like AT&T and Sprint.\textsuperscript{1}) Since networks are regulated at the national level, and involve different sets of incumbent telecom carriers in each country, there is real possibility that these global Cloud service providers can be thrust into policy issues over network operators. Moreover, the renewed interest of firms in access devices such as smart phones and tablets – until recently often considered commodity businesses that would not yield high value – make sense when considered as part of firms’ cloud offerings. Let us now use this stacks conception to map some of the major Cloud providers’ offerings and strategies.

\textbf{[Full version available by 1/2011]}